



Description of the MHS Health Level 7 Anatomic Pathology for Public Health Surveillance

Technical Document NMCPHC-EDC-TD-6-2014

By Gosia Nowak
EpiData Center Department
June 2014

Approved for public release. Distribution is unlimited.

The views expressed in this document are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government



NAVY AND MARINE CORPS PUBLIC HEALTH CENTER
PREVENTION AND PROTECTION START HERE

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188		
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</small>					
1. REPORT DATE (DD-MM-YYYY) 30-06-2014		2. REPORT TYPE Technical Document		3. DATES COVERED (From - To) June 2009 - June 2012	
4. TITLE AND SUBTITLE Description of the MHS Health Level 7 Anatomic Pathology for Public Health Surveillance		5a. CONTRACT NUMBER			
		5b. GRANT NUMBER			
		5c. PROGRAM ELEMENT NUMBER			
6. AUTHOR(S) Gosia Nowak		5d. PROJECT NUMBER			
		5e. TASK NUMBER			
		5f. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Navy and Marine Corps Public Health Center EpiData Center Department 620 John Paul Jones Circle, Suite 1100 Portsmouth, VA 23708		8. PERFORMING ORGANIZATION REPORT NUMBER NMCPHC-EDC-TD-6-2014			
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Navy and Marine Corps Public Health Center EpiData Center Department 620 John Paul Jones Circle, Suite 1100 Portsmouth, VA 23708		10. SPONSOR/MONITOR'S ACRONYM(S) NMCPHC			
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) NMCPHC-EDC-TD-6-2014			
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release. Distribution is not limited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The EpiData Center Department (EDC) at the Navy and Marine Corps Public Health Center (NMCPHC) evaluated the Health Level 7 (HL7) data source for its usefulness in health surveillance activities. This technical document provides a history of the HL7 anatomic pathology database and its contents, explains the creation of pathology records, describes the pathway of data from healthcare provider to the EDC, provides a detailed descriptions of all variables within the database, and assesses the database's strengths and limitations. Given an understanding of the strengths and limitations of the data, HL7 pathology data have proven to be a valuable source of health information for surveillance purposes. The data allow the creation of a timeline of events corresponding to a specific disease occurrence. Furthermore, data are received in a timely fashion, allowing for near-real-time surveillance of diseases.					
15. SUBJECT TERMS Health Level 7 (HL7), Anatomic Pathology, Surveillance					
16. SECURITY CLASSIFICATION OF: U			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON Gosia Nowak
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code) 757-953-0979



Abstract

The EpiData Center Department (EDC) at the Navy and Marine Corps Public Health Center (NMCPHC) evaluated the Health Level 7 (HL7) data source for its usefulness in health surveillance activities. This technical document provides a history of the HL7 anatomic pathology database and its contents, explains the creation of pathology records, describes the pathway of data from healthcare provider to the EDC, provides a detailed descriptions of all variables within the database, and assesses the database's strengths and limitations. Given an understanding of the strengths and limitations of the data, HL7 anatomic pathology data have proven to be a valuable source of health information for surveillance purposes. The data allow the creation of a timeline of events corresponding to a specific disease occurrence. Furthermore, data are received in a timely fashion, allowing for near-real-time surveillance of diseases.



Table of Contents

Abstract.....	2
Table of Contents.....	3
Executive Summary.....	5
Project Background	5
Public Health Surveillance Applications	5
Key Fields for Public Health Surveillance	5
Strengths	5
Limitations	6
Project Background.....	7
Data Origination and Flow Process.....	7
Public Health Surveillance Applications.....	8
Data Structure and Analysis.....	8
Key Fields for Public Health Surveillance	9
Defining Duplicates	9
Unique Record	9
Unique Person	9
Unique Order	9
Test Results.....	9
Corrected Records.....	10
Strengths	10
Timeliness.....	10
Completeness.....	11
Limitations.....	12
Completeness	12
Inclusion	12
Generalizability.....	12
Comparability	12
All Data Fields (Variables)	13
Automatically Populated Fields.....	13



Formatting	14
Generation of Facility Information.....	14
Field Observations (in alphabetical order).....	15
Appendix A: HL7 AP test result formats	27
Appendix B: Timeline of useful dates in HL7 anatomic pathology data	27



Executive Summary

Project Background

The EpiData Center Department (EDC) at the Navy and Marine Corps Public Health Center (NMCPHC) was tasked by the Department of Defense (DOD) Global Emerging Infections Surveillance and Response System (GEIS) with the evaluation of the Health Level 7 (HL7) data source for its usefulness in health surveillance activities. This technical document is the result of those efforts. The anatomic pathology (AP) dataset contains records documenting microscopic analysis of body tissues since 6 July 2009.

Public Health Surveillance Applications

HL7 AP data add a unique layer to the EDC's surveillance efforts. These data are not limited to physician diagnoses; therefore, they can provide laboratory testing information for tests performed among suspect cases. The greatest value of HL7 AP data for the Navy and Marine Corps currently lies in disease-specific procedures. HL7 AP testing depends on the suspect disease and may be general in the type of procedure performed (e.g. biopsy). The results of HL7 AP tests may support clinical diagnosis or treatment. The use of HL7 AP tests may be dependent on provider practice, severity of symptoms, medical history, or family history. Data on HL7 AP testing, therefore, can improve the robustness of surveillance systems based on treatment and/or International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) coded records.

Key Fields for Public Health Surveillance

Specific key fields for public health surveillance are included in the data: SPONSOR ID, FAMILY MEMBER PREFIX (FMP), SERVICE, REQUESTING FACILITY, and PERFORMING FACILITY. True duplicates are defined as records in which all fields are identical. After true duplicates are eliminated, the data can be analyzed by unique patient, test, or record. *Unique patients* are identified in the HL7 AP data through a combination of SPONSOR ID and FMP; this combination creates a unique identifier that can be used to track individual patients through all HL7 AP records. A *unique test* is defined as all records associated with each HL7 AP test. A *unique record* is defined as each record associated with each HL7 AP test for each patient.

Strengths

Several of the data fields of interest are complete but the completeness of the database as a whole continues to be assessed. Analysis showed that data were collected in the Composite Health Care System (CHCS) from the majority of the DOD military treatment facilities (MTFs). The



timeliness of data processing is within the acceptable range for the Navy disease surveillance activities, typically two days.

Limitations

It is currently not clear whether Defense Health Services Systems (DHSS) captures all CHCS HL7 AP transactions. Further work is necessary to compare HL7 AP records to other data sources to estimate completeness. The AP data only include HL7 data generated within the CHCS servers; tests performed at forward deployed, shipboard, battalion aid stations, or purchased care clinics are not captured. Incomplete demographic information (e.g., unspecified marital status, race, or ethnicity) can limit the generalizability of these data to specific minority groups. Extra precautions need to be taken when extrapolating data to larger populations and when comparing disease rates and trends among the military to non-military populations.



Project Background

The EpiData Center Department (EDC) at the Navy and Marine Corps Public Health Center (NMCPHC) was tasked by the Department of Defense (DOD) Global Emerging Infections Surveillance and Response System (GEIS) with the evaluation of the Health Level 7 (HL7) data source for its usefulness in health surveillance activities. The anatomic pathology (AP) dataset contains records documenting pathology tests performed at a military treatment facility (MTF). Records for Department of Defense (DOD) military service members (Army, Navy, Marine Corps, Air Force, Coast Guard, and US Public Health Service), overseas civilian personnel, Tricare eligible dependents, and others who receive their laboratory tests at a MTF are included in this dataset. This document describes observations on the data fields, some basic frequencies, the cleaning rules implemented for data use, and other comments relevant to the use of these data for surveillance.

Initial evaluation of the dataset involved one sample extract received by the EDC from the Defense Health Services System (DHSS). The sample extract was a very small dataset used to analyze the structure, completeness, and distribution of the entire dataset. Descriptive analysis of these data included frequency distribution of demographic fields, evaluation of null or invalid values for key fields used in surveillance, and understanding data structure in the extracts received compared to the structure of data in the Composite Health Care System (CHCS). The current data archive at NMCPHC dates back to 6 July 2009.

Data Origination and Flow Process

The HL7 AP dataset includes all anatomic pathology tests that are performed at a CHCS-based MTF. There are several mechanisms of entry. The most common process is described below, along with notable exceptions.

An HL7 AP test order is initially entered into CHCS by the ordering physician. The pathology branch within the laboratory department receives the order via CHCS and verifies it. If clarification is needed, staff may contact the ordering physician for more information. When the pathologist completes the procedure, the procedure information (e.g., test type, result text) is entered into CHCS. The record is then certified and saved on the local CHCS server. If results are edited during verification, edits are made in the CHCS record and recertified. The laboratory technician has the ability to cancel orders with physician approval. Each time a record is canceled, changed, edited, or reordered, a new record in CHCS is generated.

The HL7 AP data are limited to AP tests at MTFs that use CHCS. If orders are entered into CHCS and not completed and/or certified (test is not performed), these records do not appear in the HL7 AP dataset. Forward deployed clinics, shipboard clinics, battalion aid stations, and



purchased care facilities do not currently participate in CHCS and so tests from these facilities are not in the HL7 AP dataset.

Public Health Surveillance Applications

HL7 AP data add a unique layer to the EDC's surveillance efforts. These data are not limited to physician diagnoses or laboratory confirmed cases; therefore, they can provide supporting information for laboratory confirmed, physician diagnosed, or presumptively treated cases. The greatest value of HL7 AP data for the Navy and Marine Corps currently lies in disease-specific procedures. HL7 AP procedures depend on the suspect disease but may be general in the type of procedure performed (e.g., biopsy). The procedure type does not indicate specific disease but results may support clinical diagnosis or treatment. The use of HL7 AP procedures may be dependent on provider practice, severity of symptoms, medical history, or family history. HL7 AP data can improve the robustness of surveillance systems based on lab results and/or International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) coded records.

Current surveillance methods in the EDC include monitoring HL7 microbiology and chemistry laboratory results, ICD-9-CM codes in clinical encounter records, and outpatient/inpatient pharmacy transactions. Consequently, surveillance methods are largely disease-specific, but this specificity depends on ICD-9-CM coding practices in local MTFs, timeliness of laboratory tests, the ability to accurately flag laboratory tests of interest, and disease-specific treatment regimens. The use of HL7 AP data greatly improves the surveillance of certain diseases or conditions, such as cervical cancer, because other data on these diseases are greatly limited by laboratory test types and potential inaccuracies in ICD-9-CM coding.

Potential use of HL7 AP records is not limited to surveillance. Data on HL7 AP procedures can fill critical gaps in the military's ability to validate specific diagnoses, particularly cancer and skin conditions. Coupled with laboratory and encounter data, disease management guidelines can be evaluated. Finally, these data may provide valuable insight into clinical practice and atypical disease presentation.

Data Structure and Analysis

HL7 AP data are retrieved by the EDC in a standard, pipe-delimited flat file from DHSS via a secure connection. Each column within the data file is a fixed variable and each row should contain a unique record. Each person can have more than one distinct record, if they have multiple AP tests or updates to their tests. Each test ordered is associated with a unique record (row). The variable fields are formatted to ease analysis, except for the free text fields, which require the use of wildcards or search terms.



Key Fields for Public Health Surveillance

Defining Duplicates

Within the HL7 AP dataset there are several ways in which duplicate records can be identified. Duplicate rules described here should be checked against project objectives to ensure applicability. True duplicates are defined as records in which all fields are identical. Records meeting this criterion should be eliminated so that only one record remains. There are three types of records described here which are most relevant to public health surveillance purposes: unique record, unique person, and unique order.

Unique Record

Each record that remains after removing true duplicates is considered a unique record. There is at least one variable value different than all other records in the database.

Unique Person

Patients are identified in the HL7 AP data through a combination of SPONSOR ID and FMP. This combination creates a unique identifier that can be used to track individual patients through all HL7 AP records and across other databases. The PATIENT ID is not complete, consistent, or reliable as a source of identifying patients within or across databases. It is possible for individuals to have two separate SPONSOR IDs over time. For example, if the child of a sponsor becomes active duty, then that child will have his/her own SPONSOR ID. Each unique patient can have multiple test orders in the HL7 AP data.

Unique Order

A unique order is defined as all records associated with a single specific HL7 AP test. Each test ordered is assigned an ORDER NUMBER. ORDER NUMBERS may be reused; however, it is unlikely that a person would receive the same order number more than once. The combination of SPONSOR ID, FMP, and ORDER NUMBER can be used to identify unique orders within the HL7 AP dataset. Each unique order can have multiple records within the HL7 AP data.

Test Results

The structure of the HL7 AP data provided by DHSS was changed on 5 November 2009. This change affects how analysts use the data. Test results in HL7 AP data are in a free text field that often includes information regarding patient history, patient symptoms, provider impressions, conditions that are ruled out before final results, and final test results. This information was originally broken into segments and placed in multiple records with duplicate information for all



fields except TEST RESULT and SET ID. Records before 5 November 2009 can be sorted by unique test and SET ID to read the test results in the correct order.

Restructured data (after 5 November 2009) include all values for TEST RESULT for each unique order in the same record. This is accomplished by combining the SET ID and TEST RESULT fields into the SET ID field. The records in the new structure contain the SET ID concatenated with the missing values for the TEST RESULT field. For analyses that span the date of change, both methods of result interpretation should be applied to ensure complete case capture. An example of both record formats can be found in [Appendix A](#).

Corrected Records

The EDC currently receives records that are completed and designated with a RESULT STATUS of "F" (final). If a record is corrected (result status of "C" = amended), an additional record is generated. The information contained in the original record is included in the updated record. Additional/corrected information is appended to the SET ID RESULT TEXT field (original findings remain in this field, as well), and when the message is present, the message date/time and DHSS LOAD DATE time are updated by CHCS. If a record indicated a change is present then that record should be considered in the analysis instead of the initial record. In less than 1% of orders, the original record is corrected more than once.

Strengths

Timeliness

DHSS includes several date fields in the dataset provided to the EDC: CERTIFY DATE, COLLECTION DATE, DHSS LOAD DATE, MESSAGE DATE, ORDER EFFECTIVE DATE, and REQUESTED DATE. A timeline of useful dates is provided in [Appendix B](#). To assess the timeliness of the data, the CERTIFY DATE (date the result was certified) was compared to the MESSAGE DATE (date the HL7 message was generated by CHCS) to estimate the time between the test completion and the receipt of data at DHSS. The MESSAGE DATE was also compared to the DHSS LOAD DATE to determine the time between HL7 message generation at the local CHCS host and DHSS data parsing of the HL7 message into the database design.

For almost all records (99.8%), an HL7 message was generated the same day as the record was certified. After generation, it took approximately one day for the message to be processed by DHSS (96.8%). Based on processing of the data at DHSS, NMCPHC receives these data within approximately two days, though this time estimate needs to be verified. The brief delay in data receipt is within acceptable ranges for the Navy disease surveillance activities. Future analysis and assessment should define lag times in relation to particular MTFs or disease outcomes of interest.



Completeness

Records are received from the majority of shore-based fixed MTFs connected to CHCS, but gaps in data may exist. Gaps in data may occur due to server failure at location or due to functional errors. It is believed that HL7 AP data received by the EDC represent at least 90% of all completed HL7 AP tests in CHCS. The completeness of individual fields varies and the characteristics of each are described in detail in the [field observations](#) section. In general, some fields of particular interest, such as SPONSOR ID, FMP, and SERVICE are highly populated due to the business rules of CHCS.



Limitations

Completeness

The HL7 infrastructure at DHSS was built using pilot funds. Initially, a temporary network was created to capture HL7 messages when they were sent from CHCS host to the central server. Up until the program became formal, no back-up system existed. When the feed node fails, HL7 messages may be lost and those that have been sent may not be retrievable unless the network outage was planned for in advance. Gaps may exist in the data received at NMCPHC, though limited. Several of the identified data fields of public health interest are highly populated, but others are not. The completeness of each data field, as described below, should be considered before its use in analysis.

Inclusion

The data only includes MTFs that utilize CHCS. Forward deployed clinics, contracted managed care support clinics, and other MTFs that do not use CHCS are not captured in these data unless the order is filled by a laboratory that uses CHCS. CHCS is not used to order or process AP tests onboard ships.

Generalizability

Incomplete demographic information (e.g., unspecified MARITAL STATUS, RACE, or ETHNICITY) can limit the generalizability of these data to specific minority groups. Demographic information not provided in this database can be supplemented with other available personnel databases.

Comparability

These data are generated from the HL7 AP test records of a highly specific patient population – military service members and other military beneficiaries – which differs from the general United States (US) population in many ways, including average age, gender distribution, physical fitness, and health status. Further, this population has universal access to medical care, which is not true of the US population. These differences limit the comparability to the general US population. Extra precautions need to be taken when extrapolating data to larger populations and also when comparing the disease rates and trends of the military and non-military populations.



All Data Fields (Variables)

The following section describes frequency distributions run on all fields within the HL7 AP database, based on data through June 2012. The data fields of most interest include SPONSOR ID, FMP, SERVICE, REQUESTING FACILITY, PERFORMING FACILITY, and other fields that are necessary for the EDC's planned surveillance activities.

Automatically Populated Fields

There are several types of automatically populated fields in the HL7 AP data.

When a facility registers within CHCS, several variables are created which identify the facility: PERFORMING DMIS ID, PERFORMING FACILITY, PERFORMING FACILITY SERVICE, PERFORMING WORK CENTER, REQUESTING DMIS ID, REQUESTING FACILITY, REQUESTING FACILITY SERVICE, and REQUESTING WORK CENTER.

When DHSS compiles the data from the CHCS server, two fields are automatically populated: DHSS LOAD DATE and DHSS LOAD TIME.

Each patient or beneficiary is registered in the Defense Eligibility Enrollment Reporting System (DEERS) under the SPONSOR ID, which feeds into CHCS. When a patient presents at a medical facility, the SPONSOR ID (usually the Social Security number) is entered and their name is chosen from a drop-down list. The following patient demographic fields are automatically populated after this selection if they were entered when the patient was registered in DEERS: DATE OF BIRTH, ETHNICITY, FMP, GENDER, MARITAL STATUS, PATIENT CATEGORY, PATIENT ID, RACE, SERVICE, and SPONSOR ID. If these data are not present in the system, a designated unknown value is entered, and therefore there are no missing values in these fields. Registration is completed and records updated when the sponsor reports to a new command and selects an MTF. Administrative personnel at the MTF have the ability to edit records at the time of visit.

As records are created, edited, and completed, the date and time variables are created by CHCS system. These variables can be changed, if necessary, by the user, but this change is not common practice.

MSG DATE, MSG TIME, and MSG SENDING FACILITY are created and assigned when the message (record) is sent to the CHCS server.



Formatting

Several variables in the HL7 AP data contain numerical values. A few of these fields may contain leading zeros that would affect analysis if lost: SPONSOR ID, PATIENT ID, FMP, PERFORMING FACILITY DMIS ID, and REQUESTING FACILITY DMIS ID. To maintain data integrity, these fields should be imported in character format.

Generation of Facility Information

When each facility registers with CHCS, the facility name is created. Each record generated from the location will have the same facility name. If the facility name was entered incorrectly, it will be consistently incorrect in all records from that facility. Within each facility there are a variety of work centers that can generate HL7 AP records. The work center variable is a free text field that the ordering physician fills during order generation.

The EDC has provided DHSS with an official DOD Defense Medical Information System Identifier (DMIS ID) list. This list is used to create a four-digit DMIS ID for each record based on the information contained in the facility name field. Once records have been assigned a DMIS ID, additional fields describing the facility are created: DMIS FACILITY NAME and FACILITY SERVICE. If the DMIS ID is missing, either because the facility name was missing or a correct match was not made, these variables are also missing. Furthermore, a secondary quality assurance check is performed on the raw data once it is received at NMCPHC. Records with null values in the DMIS ID field are identified. For those records, an algorithm based on the REQUESTING and/or PERFORMING FACILITY NAME fills in the DMIS ID.

The DMIS ID is listed for both the requesting and the performing facility. REQUESTING FACILITY DMIS ID indicates which facility placed the order for the test. PERFORMING FACILITY DMIS ID indicates the facility at which the test was performed.



Field Observations (in alphabetical order)

ACCESSION NUMBER

The format of the ACCESSION NUMBER is a combination of the 1) date in an YYYYMMDD format, 2) a two or three character alpha setting, and 3) a numeric listing of how many tests of that specific type were run in one day. The last numeric digits can range from 1 to 9999. ACCESSION NUMBERS are created for each unique biological sample collected from the patient. Different HL7 AP tests from the same biological sample can have the same ACCESSION NUMBER. These numbers could be recycled throughout a day's time, and should not solely be used to identify a record. They may be used to determine tests ordered per patient in conjunction with the SPONSOR ID, FMP, and the date when the test was ordered. There are missing values in less than 1% of records.

BODYSITE COLLECTIONSAMPLE

The BODYSITE COLLECTION SAMPLE refers to the place on the body where the specimen is collected from the patient. This field is used with SPECIMEN SOURCE to determine where the sample is taken. A patient can have numerous samples taken from one area (i.e. a lung can have numerous biopsy specimens, thus having a different ACCESSION NUMBER for each specimen). But, like SPECIMEN SOURCE, it can be used to determine if proper protocol was used for a test, or can be used to determine the type of test performed (i.e. PAP smear would not have a non-cervical sample type). BODYSITE COLLECTION SAMPLE is missing in 38% of HL7 AP records.

CERTIFY DATE

The CERTIFY DATE is the date when a laboratory technician certifies the results into CHCS, or makes changes within the system. Unlike the ORDER EFFECTIVE DATE, there can be deviations between the values within SET ID, due to different test run dates. The CERTIFY DATE is formatted YYYYMMDD and contains limited missing values. There are less than 1% of records missing a value in this field. The values of the timeframe are between ORDER EFFECTIVE DATE and MSG DATE.

CERTIFY TIME

This field represents the time component of the CERTIFY DATE and is formatted using a standard 24 hour clock. The possible values are from 0000 to 2359. There are less than 1% of records missing a value in this field.



CLINICAL COMMENTS

The CLINICAL COMMENTS is a free text field which allows the provider or laboratory technician to add additional information regarding the patient's symptoms, quality assurance testing information, contact phone numbers, specimen media, or instructions on test procedures. All records in the HL7 AP database have missing values for this field.

This field is not primarily used in case definition, but in other databases it is added to eliminate superfluous entries.

COLLECTION DATE

The COLLECTION DATE is the date when the specimen is extracted from the patient. The value for this entry should be between the values of the ORDER EFFECTIVE DATE and the CERTIFY DATE. The COLLECTION DATE is formatted YYYYMMDD and there are no missing values.

Since the field approximates the day that the laboratory sample is collected, it may be useful for analysis. It can be used for time analysis between the specimen collection and test result certification. By knowing the timeframe of each test conducted, an approximation of the type of test used can be determined.

COLLECTION TIME

As with COLLECTION DATE, the COLLECTION TIME is the time when the specimen is extracted from the patient, and follows a standard 24-hour clock. Unlike ORDER EFFECTIVE TIME, the timeframe is from 0001 to 2400. All times are valid entries. There are no missing values.

CPT CODE DATA

The CPT CODE DATA is an alphanumeric field which identifies a particular test by the Current Procedural Terminology (CPT) code. The CPT code is defined by the American Medical Association, and describes medical, surgical, and diagnostic procedures. This is designed to communicate uniform information about medical services and procedures between physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytical purposes.

The variable format is #####\##\AD. The first group of characters defines the CPT code used within the HL7 AP dataset. The second portion is a modifier code which indicates the accession area and work element. There are multiple codes listed in the CPT codebook. The values observed in the HL7 AP data are defined as: 26 – Professional/Pathologist, 32 – Mandated



Service (MTF performs laboratory for a branch clinic), 90 – Reference Laboratory Service (e.g. LabCorp), or 91 – Repeat Clinical Diagnostic Procedure (multiple tests for subsequent results). The value of 00 is present but is undefined by the reference.

The regional CHCS site maps a CPT code to a particular methodology or technique. CPT codes are assigned at various levels to CHCS test files when the laboratory sets up the procedure. All tests that do not have a specific CPT code may be given unlisted procedure/service codes defined for the specific types of test (immunology, chemistry, microbiology, hematology, etc.).

Values are missing in this field in 35% of HL7 AP records.

DATE OF BIRTH

The DATE OF BIRTH field (DOB) is included in the format YYYYMMDD. It is possible to have inaccurate values for DOB. If the full DOB is unknown but the year of birth is known, then CHCS automatically enters zeros for the month and day. Less than one percent of records have either missing a month and day or are completely missing the date of birth.

DHSS LOAD DATE

DHSS LOAD DATE indicates the date when DHSS loads the data from the central CHCS server. When present, this field could be used to determine the timeliness of reporting and to identify lags in reporting times from certain MTFs. The format is YYYYMMDD. Though this field should be automatically generated, the value for this field is missing in 99% of HL7 AP records.

DHSS LOAD TIME

Time component of the DHSS LOAD DATE field, and is formatted: HHMM. The values present in the data are 0300, 1000, 1600, and 2000. The value for this field is missing in 99% of HL7 AP records.

ETHNICITY

ETHNICITY is an alphanumeric field with six possible values; 1=Hispanic, 2=South Eastern Asian, 3= Filipino, 4=Other Asian Pacific Islander, 9=Other, and Z=Unknown. There are no records missing a value in this field. The most frequent group is Unknown with 51%, and 43% of records in the HL7 AP database are categorized as Other. These results indicate that the field of ETHNICITY may be self-identified and not consistently reported. Those entries which are not reported are labeled as Unknown. The Unknown responses are assumed to be pre-populated in order to eliminate blanks within the database. The number of Unknown or Other responses



limits the ability to identify disease trends in minority groups and to identify diseases that have a disproportionate burden on these groups.

FMP

FMP is the family member prefix that designates the relationship of the patient to the sponsor. The distribution of FMP among the records is as expected, with most frequent values of 1-3, 20, and 30 which are values that correspond to first, second, and third child of sponsor (FMP=1-3), the sponsor (FMP=20), and spouse of sponsor (FMP=30). All entries have a value for FMP.

GENDER

There are three values possible for the GENDER field; M=Male, F=Female, X=Unknown. There are no records with a missing value in this field, and less than 1% are coded as Unknown.

MARITAL STATUS

There are nine values for MARITAL STATUS: A=Annulled, D=Divorced, I=Interlocutory Decree, L=Legally Separated, M=Married, S=Single/Not Married, W=Widow or Widower, Z=Unknown. There are no missing values for records in the HL7 AP dataset. The majority of records are classified as Unknown (47%). The next highest group is Married (42% of records) followed by Single/Not Married (9% of records).

MEPRS CODE

The MEPRS CODE is a four alphanumeric code that indicates the location within the MTF the person is seen. The first letter indicates the most general area and translates as: A=inpatient, B=outpatient, C=Dental, D=ancillary, E=support services, F=special programs, and G=medical readiness. It is advised to obtain an up-to-date list of all possible codes. The HL7 AP dataset does not have missing values because it is automatically populated when the record is created. This field is useful for tracking where people are seen within the MTF (ambulatory care, special dialysis clinics, the maternity ward, etc.) which can affect the interpretation of the data. The majority of records present in the HL7 AP dataset have a MEPRS code that begins with B (90%).

MSG DATE

This field is formatted YYYYMMDD. There are no missing values and all are valid dates. This date approximates the transaction time between the MTF and the regional CHCS site, but it can vary based on location. Some MTFs send messages in batches, therefore the time or date portions may not correlate to the actual transaction time.

MSG ID



The Message ID (MSG ID) is an alphanumeric code assigned to each batch of messages based on when the message is sent from CHCS to the server. The MSG ID is not unique to each record; each batch of messages is assigned one MSG ID. The MSG ID format varies by MTF and may include numbers, letters, or numeric code that identifies the MTF, or it can identify the function of the message (i.e. RESCHED-057342).

MSG SENDING FACILITY

This field is formatted as AA###. This field allows analysts to identify and track the transfer of messages from the MTFs to DHSS and the EDC. There are missing values in less than 1% of records within this dataset.

MSG TIME

The MSG TIME is the time when the message is sent from the MTF to the regional CHCS site, and follows a standard 24-hour clock. The numbers range from 0001 to 2359. There are no recorded times for 0 or 2400. All times are valid entries. There are no missing values.

NO OF CPT CODES

The NO OF CPT CODES is a numeric field which lists the number of CPT codes used for each test performed. The number of CPT codes is determined at each regional location, and is missing in 35% of records. This field is currently not used within the EDC.

ORDER EFFECTIVE DATE

The ORDER EFFECTIVE DATE is the date that the laboratory order enters CHCS. It is different from the MSG DATE since the MSG DATE is generated after the laboratory results are certified. The ORDER EFFECTIVE DATE more accurately approximates when the laboratory test is actually ordered. The ORDER EFFECTIVE DATE is formatted YYYYMMDD and less than 1% of values are missing. Since the field approximates the time that the laboratory test is ordered, it may be useful for analysis. It could be used to identify when the patient presented with clinical symptoms necessitating the test, to allow for time analysis between the order dates and sample collection date, to assist in determining a duration until the completion of the test, to determine which type of test is used, and to identify time lags between when the test is ordered and when data is available for analysis at the EDC.

ORDER EFFECTIVE TIME

This field represents the time component of the ORDER EFFECTIVE DATE and is formatted using a standard 24-hour clock. Unlike MSG TIME, this timeframe includes values for 0000. The range present is 0000 to 2359, and less than 1% of values are missing.



ORDER NOTES COMMENTS

The ORDER NOTES COMMENTS is a text field which allows the provider to include notes or comments that accompany the test ordered. This field is not currently populated in the dataset.

ORDER NUMBER

The ORDER NUMBER is a numerical code with eleven digits (xxxxxx-xxxxx) unique to each order but not unique for each record. These numbers are unique for each location, and are not circulated. The first set of numbers is the date, and the last five numbers are consecutive for tests provided at that specific location. An order can have multiple records that correspond to changes made to the order (i.e. changes in test, cancellations). All changes appear as individual records with the same ORDER NUMBER. It is a plausible way to track a patient but it is not useful for identifying unique records.

ORDERING PROVIDER

The ORDERING PROVIDER field indicates the name of the ordering physician. It has three components each separated by a comma: Last Name, First Name, and Middle Initial. It is structured to facilitate analysis but could be separated if necessary. Values are missing for this variable in 40% of records.

PATCAT CODE

The patient category code (PATCAT CODE) is an alphanumeric code that indicates the patient's status with the uniformed services. The first letter of the code refers to the branch of service of the sponsor (A=Army, B=National Oceanic and Atmospheric Administration, C=Coast Guard, F=Air Force, K=other beneficiaries of the federal government, M=Marine Corps, N=Navy, P=US Public Health Service, R=NATO recipient). It is followed by two digits corresponding to the patient relationship to the sponsor. For example: A11=Army Active Duty Member, A41=Army Dependents of Active Duty, etc. A complete list should be obtained from DOD resources. Less than 1% of records are missing PATCAT CODES in the HL7 AP database.

PATIENT ID

The PATIENT ID is intended to serve as a unique identifier for each patient. The format for PATIENT ID is a nine digit numeric listing. The PATIENT ID is the patient's SSN when available. PATIENT ID is missing in less than 1% of records. The value of PATIENT ID cannot be validated based on the data received by the EDC. The SPONSOR ID in conjunction



with FMP should be used as a substitute unique patient identifier. Importing this field in character format can prevent the loss of leading zeros.

PERFORMING DMIS FACILITY NAME

This field is the text translation of the DMIS ID provided in the PERFORMING DMIS ID field. This field is assigned by DHSS at the request of the EDC. The translation of the DMIS code on the official list is often more accurate than the PERFORMING LOCATION FACILITY field in CHCS. Use of this field allows for more accurate analysis of geographic information. Since the field is also a translation of the PERFORMING LOCATION FACILITY field in CHCS, it will be missing when that variable has a missing value (4% of records).

PERFORMING DMIS ID

The PERFORMING DMIS ID is a four digit code assigned by the DOD to all units at all installations to uniquely identify them. The EDC provided an official DMIS list to DHSS for the purpose of creating this variable. DHSS translates the PERFORMING LOCATION FACILITY field within CHCS to its assigned DMIS code. This code allows for grouping of MTFs based on geographic location, as well as identification of parent/child relationships between installations. Since this field is calculated based on the PERFORMING LOCATION FACILITY field, all records missing a value for that field will be missing a value for the PERFORMING DMIS ID field (4% of records). Importing this field in character format can prevent the loss of leading zeros, which may produce complications when producing summary statistics.

PERFORMING FACILITY SERVICE

The PERFORMING FACILITY SERVICE field indicates the branch of service with which the MTF is associated. This value is determined from the DMIS code list provided to DHSS by the EDC. It is missing when the Performing Facility information is missing (4% of records). The possible values are: A=Army, F=Air Force, and N=Navy. This field is useful for limiting the observations included in any investigation. Often, the data available for use are limited by branch of service for the MTF or patient. If this is the case, the HL7 AP data can be limited to the same parameters.

PERFORMING LOCATION FACILITY

The performing facility field in CHCS indicates the name of the MTF where the test was performed. Problems are encountered if the text is entered incorrectly when the facility is registered in the system (i.e. misspellings). Values in this field are missing in 4% of records.

PERFORMING LOCATION WORK CENTER



The PERFORMING LOCATION WORK CENTER field indicates the work center within the laboratory that provided the service. This field is an unstructured text field with many possible values.

RACE

There are six possible values for RACE: C=White, M=Asian or Pacific Islander, N=Black, R=American Indian or Alaskan Native, X=Other, and Z=Unknown. There are no records missing a value for RACE; however, 47% of the records are classified as Unknown. The Unknown responses are assumed to be pre-populated, to eliminate blanks within the database. This limits the ability to use the data to look at diseases or conditions that disproportionately affect one race.

RECORD TYPE

The value "LAP" for RECORD TYPE identifies the HL7 AP dataset. All entries in this dataset have the value of LAP in this field.

REQUESTED DATE

The REQUESTED DATE is a date field formatted as YYYYMMDD, and there are no missing values. This field is not frequently used in data analysis, as a detailed definition is not available.

REQUESTED TIME

This field represents the time component of the REQUESTED DATE formatted using a standard 24-hour clock. The timeframe is from 0000 to 2359, and there are no missing values. This field is not frequently used within the time analysis, as the ICD does not provide a detailed definition.

REQUESTING DMIS FACILITY NAME

This field is the text translation of the DMIS ID provided in the REQUESTING DMIS ID field. This allows for more accurate investigations when geographic information is used, because it is created using an official DOD DMIS list. Because this field is a translation of the REQUESTING DMIS ID field in CHCS, it will be missing when that field is missing in the record (6% of records).

REQUESTING DMIS ID

The REQUESTING DMIS ID is a four digit code assigned by the DOD to all units at all installations to uniquely identify them. The EDC provided an official DMIS list to DHSS for the purpose of creating this variable. DHSS translated the PERFORMING DMIS FACILITY



NAME field within CHCS to its assigned DMIS code. This code allows for grouping of MTFs based on geographic location, as well as to identify parent/child relationships between installations. Since this field is calculated based on the PERFORMING DMIS FACILITY NAME field, all records missing a value for that field will be missing a value for the PERFORMING DMIS ID field (6% of records). Importing this field in character format can prevent the loss of leading zeros, which may produce complications when producing summary statistics.

REQUESTING FACILITY NAME

The REQUESTING FACILITY NAME is the field in CHCS that indicates the name of the MTF where the order originated, and is a relatively standard text field. Problems are encountered if the text is entered incorrectly when the facility is registered in the system (i.e. misspellings). The field allows tracking of orders from origin to where they are filled. Values are missing in this field for 44% of records, so REQUESTING DMIS ID or REQUESTING DMIS FACILITY NAME should be used for location identification purposes in HL7 AP data.

REQUESTING FACILITY SERVICE

The REQUESTING FACILITY SERVICE field indicates the branch of service with which the MTF is associated. This value is determined from the DMIS code list provided to DHSS by the EDC. It is missing when the performing facility information is missing (6% of records). The possible values are: A=Army, F=Air Force, and N=Navy. This field is useful for limiting the observations included in any investigation. Often, the data available for use are limited by branch of service for the MTF or patient. If this is the case, the HL7 AP data can be limited to the same parameters.

REQUESTING WORK CENTER NAME

The REQUESTING WORK CENTER NAME is the ward or clinic within the MTF that requests the laboratory test. This field is an unstructured text field with many possible values. Values are missing in less than 1% of records.

RESULT NOTES

The RESULT NOTES field is a character string which allows the laboratory technician to provide additional information about the result, a recommendation for additional testing, or the interpretation of the laboratory result. This field is not populated in the HL7 AP database.

RESULT STATUS OBX



The RESULT STATUS OBX field is a character string which shows the status of the test performed. There are three entries which are used: P (Preliminary), F (Final), and C (Correction). These tests are used in a timely fashion, and always follow the order: P, F, C. A test always has an F within a SET ID of a test, but may also have a P or a C. Should a test have more than one RESULT STATUS OBX, it has the same SET ID, TEST NAME, and TEST ORDERED, but is on a separate entry line. An entry of "C" is entered when the record is amended due to operator error, wrong test ordered, the test is performed for the wrong patient, or test results need to be updated for any other reason. There are no missing values for this variable.

SERVICE

The service field refers to the service branch of the sponsor. The value is determined from the first component of the PATCAT field. There are missing values for this variable in less than 1% of records. The highest proportion of records belonged to the Army, Navy/Marine Corps, and Air Force, respectively.

SET ID

The SET ID field was affected by the DHSS restructure effective 5 November 2009. Original structure records contain the SET ID only, while restructured records include a concatenation of the SET ID and TEST RESULT fields.

Prior to 5 November 2009

The HL7 AP test results are a free text field divided into multiple records, and values for all other variables in the records are the same. The SET ID allows analysts to order the records correctly to review the full results. The SET ID variable is a numeric field used to identify the logical order of test results within an HL7 message. There are missing values in less than 1% of records.

After 5 November 2009

Restructured data include all values for TEST RESULT for each unique order in the same record. This is accomplished by combining the SET ID and TEST RESULT fields into the SET ID field. The records in the new structure contain the SET ID concatenated with values for the TEST RESULT field.

SPECIMEN SOURCE

The SPECIMEN SOURCE is a text field which describes the type of specimen tested. This field is useful to determine if the proper protocol is used for a laboratory test.



This field is used with BODY SITE COLLECTION SAMPLE to determine where the sample is taken. A patient can have numerous samples taken from one area (i.e. a lung can have numerous biopsy specimens, thus having a different ACCESSION NUMBER for each specimen). But, like BODY SITE COLLECTION SAMPLE, it can be used to determine if proper protocol was used for a test, or can be used to determine the type of test performed (i.e. PAP smear would not have a non-cervical sample type). SPECIMEN SOURCE is missing in less than 1% of HL7 AP records, but can include a value of "NULL".

SPONSOR ID

The SPONSOR ID field corresponds to the SSN of the sponsor and is formatted xxxxxxxx with no dashes.

The SPONSOR ID is not sufficient to serve as a unique identifier for each patient, but it can be used in conjunction with the FMP to create a unique patient identifier. It is important to preserve the entire SSN when importing the data into any analysis program. If the field is not properly coded as a character field, leading zeros will be dropped.

Not all SPONSOR IDs are Social Security Administration SSNs. If the patient does not have a valid SSN, a pseudo SSN is created. The pseudo Sponsor ID begins with 800 or 900, followed by the date. If the number is already assigned to another patient, the primary three numbers will change to 801 or 901 consecutively depending on the number created with the same date.

Additionally, quality assurance testing is conducted in laboratories. Quality assurance procedures utilize SSN-like identifiers in the SPONSOR ID field. The Sponsor ID for these procedures may resemble a pseudo-SSN, arbitrary identifiers such as 777777777, or three consecutive zeros. These tests will have labels such as Ztest, Quality Control, PSR, CAP, Non-human (NH,#), etc.

TEST NAME

The TEST NAME is a text field that shows which test is performed on the sample provided. This value is usually generated from a drop-down list of tests related to the TEST ORDERED variable. The variance between test names suggests the fields are automated at the regional CHCS level. The TEST NAME includes entries such as tests to be performed, quality controls, temperature, and even alerts for positive results. Quality control tests are within this field, and are noted via a ZZZ prior to the actual test name. There are missing values in less than 1% of records.

TEST ORDERED

TEST ORDERED identifies the requested observation, test, or panel. Each regional CHCS location has the autonomy to determine the criteria for each test ordered. Therefore, the TEST



ORDERED field can have different groupings of tests per DMIS location. The TEST ORDERED value is repeated among all records for tests associated with it according to the ORDER NUMBER. A provider can use a drop-down menu to determine the test(s) to be performed on a specimen. This shows all available tests per each test ordered. There are no missing values for this variable.

TEST RESULT

The TEST RESULT is an alphanumeric field which shows either the pending information or the final results of a test ordered. There are multiple variations, including misspellings and slang language (i.e. NOPERS). TEST RESULT is missing in 28% of records. In the HL7 AP dataset, test results can be found in the SET ID field after 5 November 2009. Protected Health Information (PHI) has been identified within the test results and caution should be used when removing personal identifiers within data to include this field.



Appendix A: HL7 AP test result formats

Original Record Format (before XXXXX):

Set Id	Test Name	Test Result
7	Tissue Exam	RECEIVED ARE TWO CEREBRIFORM TAN OVOIDS OF SOFT TISSUE EACH MEASURING
8	Tissue Exam	APPROXIMATELY 4.0 X 2.5 X 1.5 CM. THE EXTERNAL SURFACE OF EACH IS UNREMARKABLE. THE RIGHT
9	Tissue Exam	ONE IS SAMPLED IN CASSETTE A1 AND THE LEFT IN A2. 2SS. 21 DEC 09 FINAL DIAGNOSIS: TONSILS, BILATERAL,
10	Tissue Exam	TONSILLECTOMY: FOLLICULAR HYPERPLASIA.

Revised Record Format:

TEST NAME	SET ID RESULT TEXT
Tissue Exam	~7, TISSUE EXAM, RECEIVED ARE TWO CEREBRIFORM TAN OVOIDS OF SOFT TISSUE EACH MEASURING ~8, TISSUE EXAM, APPROXIMATELY 4.0 X 2.5 X 1.5 CM. THE EXTERNAL SURFACE OF EACH IS UNREMARKABLE. THE RIGHT ~9, TISSUE EXAM, ONE IS SAMPLED IN CASSETTE A1 AND THE LEFT IN A2. 2SS. 21 DEC 09 FINAL DIAGNOSIS: TONSILS, BILATERAL, ~10, TISSUE EXAM, TONSILLECTOMY: FOLLICULAR HYPERPLASIA.

Appendix B: Timeline of useful dates in HL7 anatomic pathology data

